Breakthrough Efficiency in NLP Model Deployment

Marshall Choy is the Vice President of Product at SambaNova Systems, responsible for product management and go-to-market.

Throughout their lifecycles, modern industrial NLP models follow a cadence. They start from one-time task-agnostic pre-training and then go through task-specific training on quickly changing user data. These periodically updated models are eventually deployed to serve massive online inference requests from applications.

A current active research trend is deploying state-of-the-art NLP models, like BERT, for online inference. As models grow larger each year, there is growing debate on how to deploy these models in real-time pipelines. To enable practical deployment, various techniques have been developed to distill large models down to compact variants. In applications such as digital assistants and search engines, these compact models are the key to attaining low-latency, high-accuracy models that satisfy service level requirements.

SambaNova Systems provides a solution for exploring and deploying these compact models—from a single SambaNova Systems Reconfigurable Dataflow Unit (RDU) scale to multiple SambaNova DataScale systems scale—delivering unprecedented advantages over conventional accelerators for low-latency, high-accuracy online inference.

The Proven Power of Dataflow Execution on RDU

The latency of compact models on GPU are fundamentally limited by its kernel-based execution mode. For online inference with batch size 1, the overhead of context switching and off-chip weight memory access for operation kernels can dominate latency on traditional architecture. SambaNova RDU is built on the SambaNova Systems Reconfigurable Dataflow Architecture (RDA) to remove this barrier. Specifically, on a recently proposed compact BERT model, TinyBERT, the RDU can attain **5.8X** latency speedup over V100 GPU for MNLI, a popular text classification task.

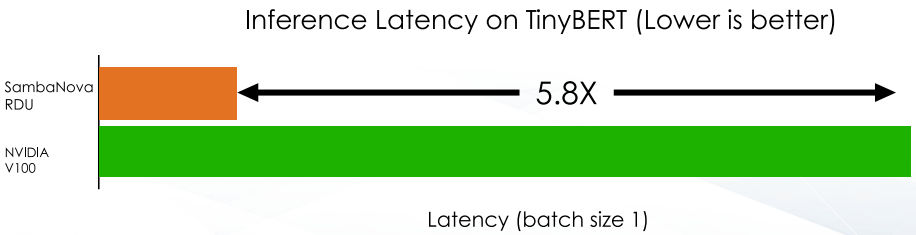


Fig 1: Latency comparison for online inference

In applications such as a digital assistant or a search engine, the input data are natural language tokens with short sequence length, e.g., smartphone assistant queries such as "What is the weather in San Francisco?". For these types of scenarios, reduced sequence length typically has a negligible impact on the accuracy attained by compact models. This is another characteristic that is deeply coupled with the latency advantage of RDUs. While a GPUs latency saturates with reduced sequence length for compact models, the RDUs latency improves with reduced sequence length.

As shown in Figure 2, the TinyBERT model can match state-of-the-art model accuracies across sequence lengths from 64 to 256 on the MNLI benchmark task that we use as a proxy. In Figure 3, we can see that the GPU demonstrates the same latency across sequence lengths. However, the speedup of RDU over GPU is boosted to **8.7X** at reduced sequence length of 64.

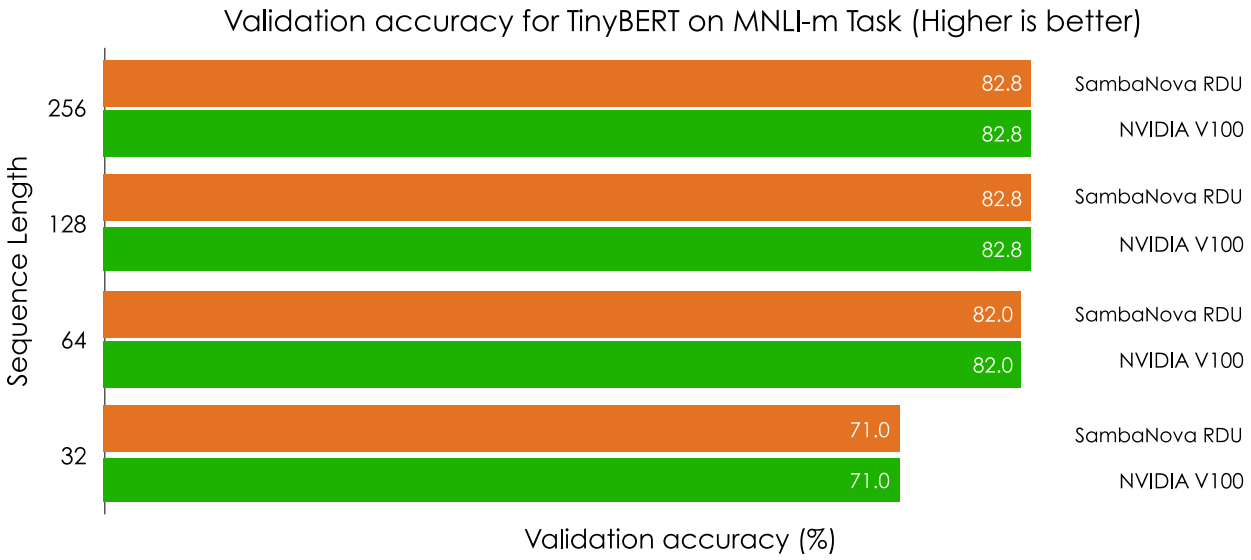


Fig 2: RDU and GPU model accuracy for different sequence length

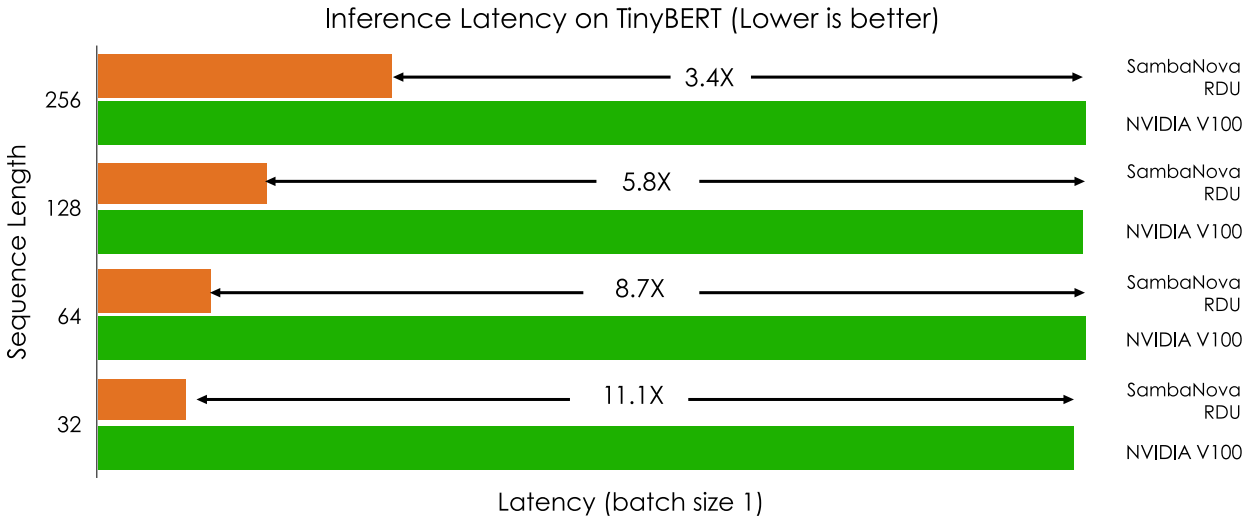


Fig 3: Bar chart for RDU and GPU latency for different sequence length

Amplifying Accuracy With SambaNova Systems DataScale

Our dataflow-optimized chip demonstrates unprecedented capability for low-latency online inference for compact models. Utilizing these capabilities from the dataflow chip our research labs have also shown the full SambaNova DataScale systems (8-sockets) can be used to attain bleeding-edge accuracy while performing low-latency inference on compact NLP models.

The study from the SambaNova Systems research lab shows that majority voting across multiple model instances can significantly boost the accuracy attained by the TinyBERT (Fig. 4). The SambaNova DataScale system is perfectly designed to efficiently exploit these accuracy gains. We show that we can deploy multiple TinyBERT models on to all eight sockets of the SambaNova DataScale system. As shown in Fig 5, when ensembling TinyBERT models, the classification accuracy is boosted for 0.4% at negligible cost on latency compared to a single TinyBERT model on an RDU.

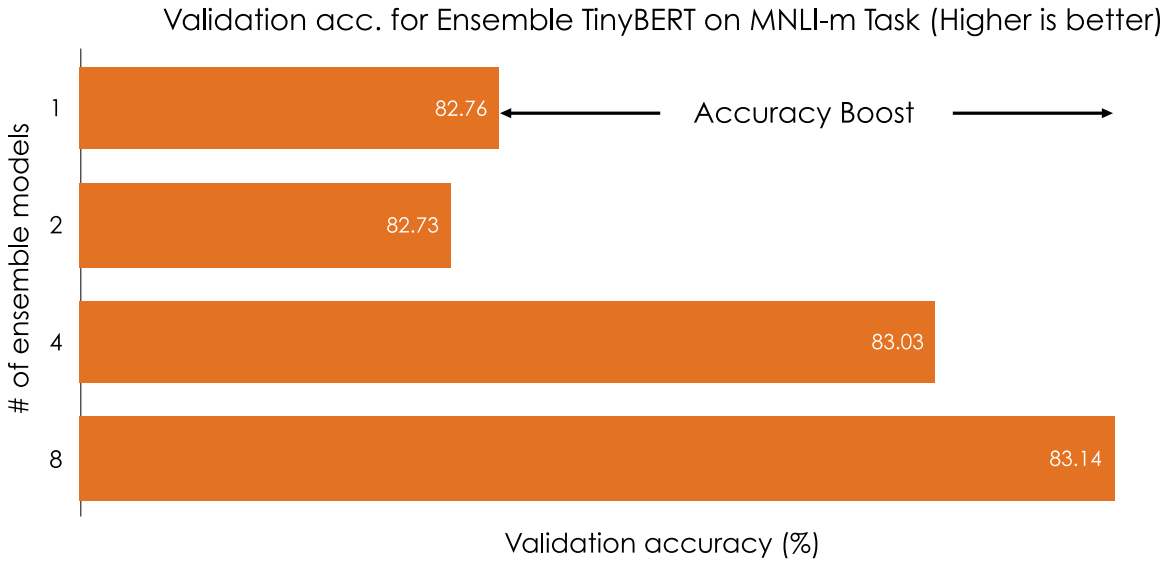
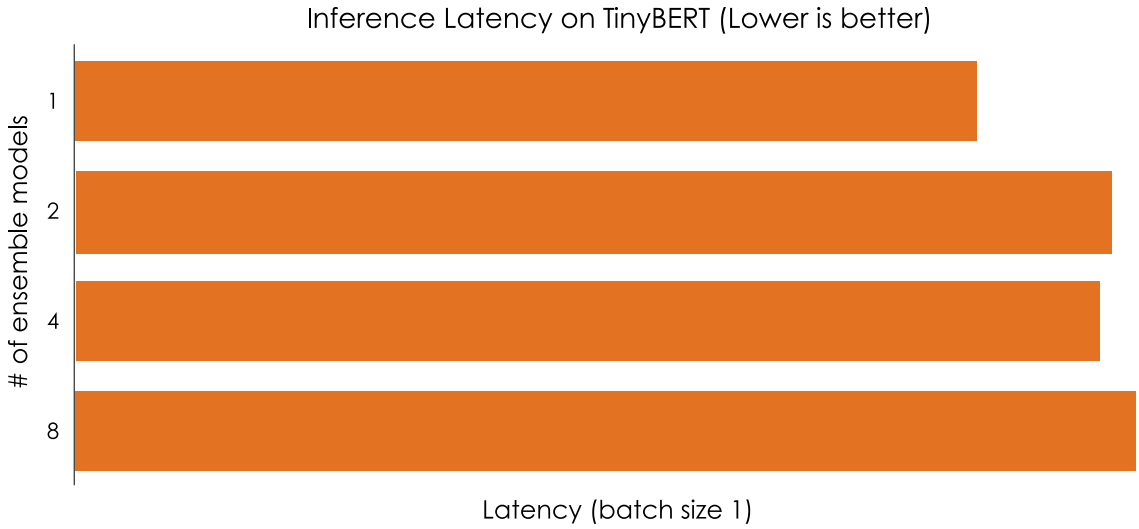


Fig 4. Model accuracy with different numbers of experts for ensemble



The compact BERT model is just one important case where our SambaNova Systems DataScale provides a tailored solution for low-latency, high-accuracy online inference.

Fig 5. Comparison of latency for single TinyBert on one RDU and 8 experts on 8-socket systems

SambaNova provides state-of-the-art technologies to support NLP, high-resolution computer vision, and recommender models. To learn more, request a meeting.

Request a Meeting

sambanova.ai